

**SYSTEM AND METHOD FOR MANAGING
A TEXTUAL ARCHIVE USING SEMANTIC UNITS**

ABSTRACT OF THE DISCLOSURE

A system and method for indexing and searching textual
archives using semantic units such as syllables and
morphemes. In one aspect, a system for indexing a textual
archive comprises an AHR (automatic handwriting recognition)
system and/or OCR (optical character recognition) system for
transcribing (decoding) textual input data (handwritten or
typed text) into a string of semantic units (e.g., syllables
or morphemes) using a statistical language model and
vocabulary based on semantic units (such as syllables or
morphemes). The string of semantic units that result from a
decoding process are stored in a semantic unit database and
indexed with pointers to the corresponding textual data in
the textual archive. In another aspect, a system for
searching a textual archive is provided, wherein a word (or
words) to be searched is rendered into a string of semantic
units (e.g., syllables or morphemes) depending on the
application. A search engine then compares the string of
semantic units (resulting from the input query) against the
decoded semantic unit database, and then identifies textual
data stored in the textual archive using the indexes that
were generated during a semantic unit-based indexing
process.

YOR9-2000-0144US1 (8728-375)